Hybrid_contig_labelling_summary

October 11, 2022

1 Hybrid assemblies analysis

`Aniket Mane & Cedric Chauve`
`2022-10-11`

1.1 Overview

This document contains a very preliminary analysis of the the *Enterococcus* data sets. It focuses on two questions: - accuracy of using the hybrid assemblies to define ground truth plasmids; - performances of plasmid prediction methods on the data sets.

The main outcome of the first part of the analysis is that the hybrid assembliers do not provide an accurate source for esablishing plasmids ground truth, thus rending the second part much less informative.

1.2 Data

The statistics about hybrid assemblies are available in the files `ctg_details_Efaecalis.csv` and `ctg_details_Efaecium.csv`.

The results of **HyAsP**, **MOBSuite** and **PlasBin** are available in the files `isolate_details_Efaecalis.csv` and `isolate_details_Efaecium.csv`.

1.3 Labelling hybrid contigs

The goal of labelling hybrid contigs is to identify which such contigs are likely plasmid contigs, in order they can be used to define the ground truth for the short-reads assemblies, through mapping short-reads contigs onto hybrid contigs.

For each species, we determine, from the length distribution of the circular hybrid contigs for all isolates, a **threshold L** that separates short circular contigs (plasmids) from other contigs (chromosomes).

Then we assign a class to each hybrid contig as follows: - circular contigs of length at most **L** are labelled as *plasmids*; - contigs (circular or otherwise) that are longer than **L** are labelled as *chromosomes*; - all remaining contigs, i.e. linear contigs of length below **L**, are labelled as *ambiguous*.

This has the downside that if a plasmid has not been fully assembled, and appears in a set of short linear contigs, then they will all be labelled as ambiguous, and this will impact how we evaluate plasmid predicion methods.

In the experiments below, we look at the prevalence of such ambiguous contigs.

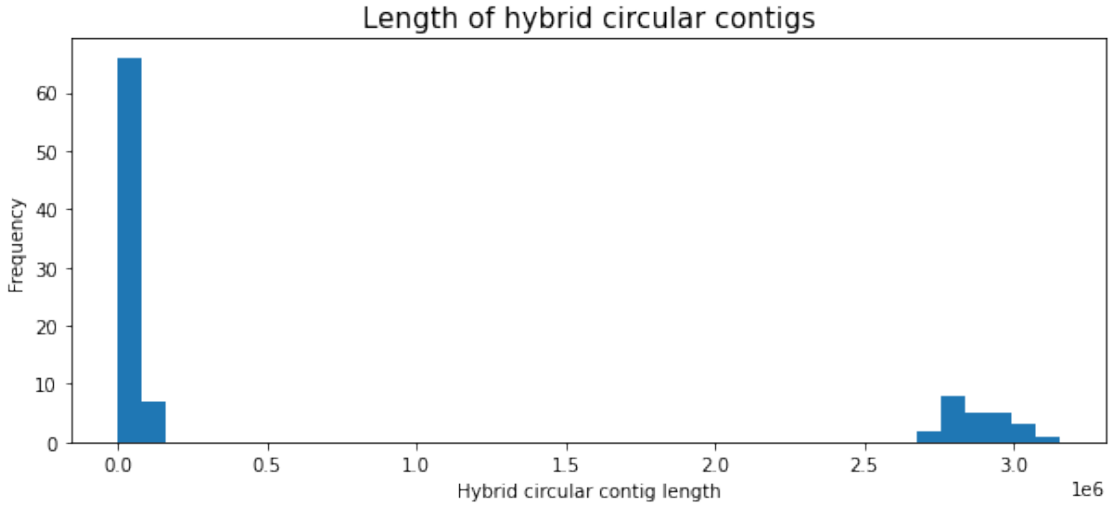
1.4 E. faecalis

We present our analysis of the *E. faecalis* isolates. In the following, contig refer to hybrid contigs.

Statistics on isolates and contigs

```
The total number of E. faecalis isolates is 60 and of contigs over all isolates
is 958
```

Length distribution of circular contigs The key parameter to labelling contigs as is the threshold **L**, that has to be chosen in an informed way. Our approach is to consider only circular contigs and to select a threshold that separates long circular contigs (likely from chromosomes) from smaller ones (likely plasmids).

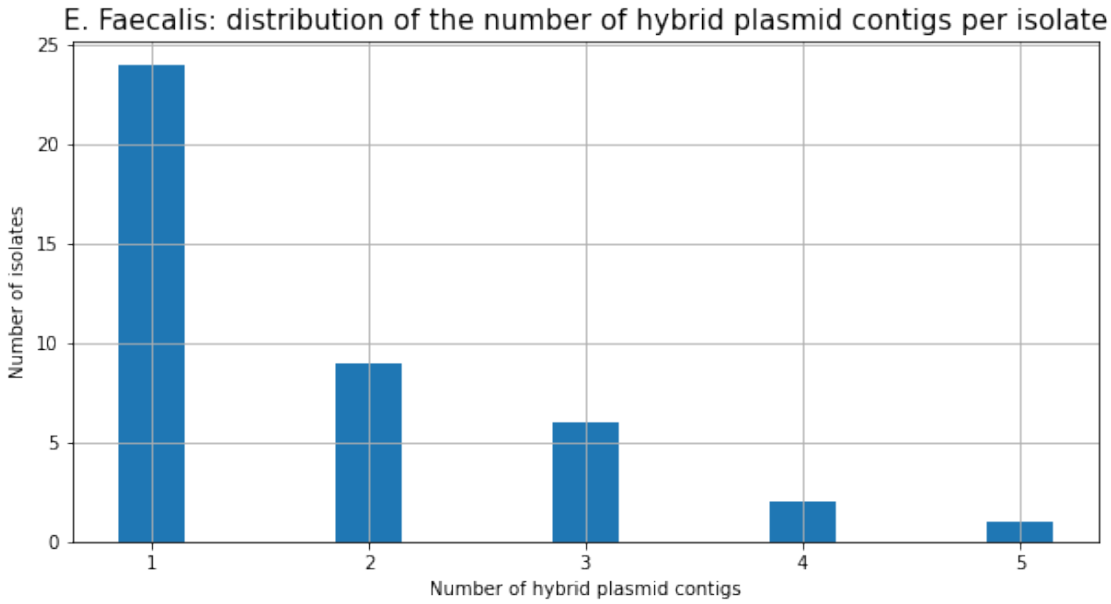


The figure above shows a very clear separation between short and long circular contigs, however leaving a very large gap where to chose **L**. We take a stringent approach and choose a small value of $\mathbf{L = 300000}$.

```
Over all E. faecalis isolates there are 97 circular contigs
  24 are classified as chromosome
  73 are classified as plasmid
```

The small number of chromosome contigs compared to the total number of isolates shows that on average the chromosome is not fully assembled into a circular contig.

Hybrid contigs classified as plasmids



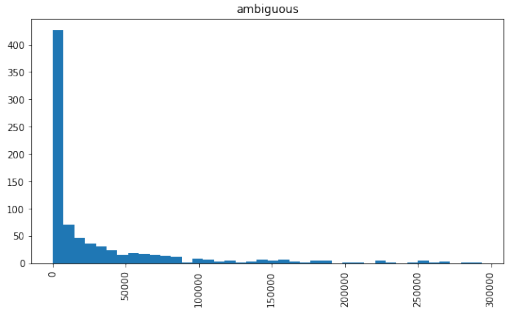
The above figure seems reasonable. We do not see isolates with an unexpectedly large number of plasmids (reminder that as we impose circularity to contigs labelled plasmids, we expect each such contig to be a fully assmebled plasmid).

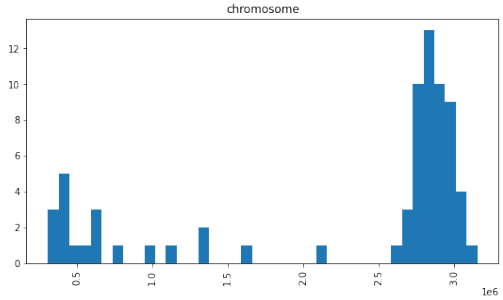
Distribution of contig lengths per label We now look at the three classes of contigs (plasmid, chromosome and ambiguous), especially to understand better features of ambiguous contigs, whose large number casts a shadow on the whole project.

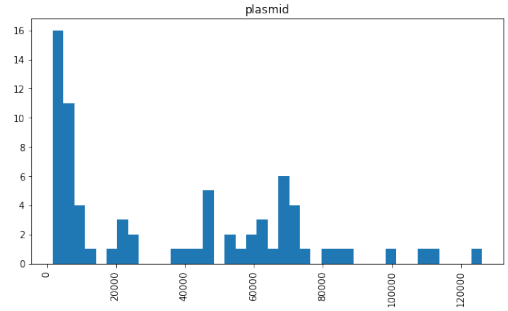
```
Number of chromosome contigs: 71
Number of plasmid contigs: 73
Number of ambiguous contigs: 814
```

Overall we see that a very large number of contigs are ambiguous. We will discuss this later. But first we look at the length distribution of the contigs per class.

```
Hybrid contig length distribution per class
```







`887 out of 958 contigs are under 300000 bp in length and 814 of these contigs have been classified as ambiguous because they are linear.`

Most of the ambiguous contigs are very short. There is a strong separation between plasmid and chromosome contigs, with the chosen threshold (300000bp) to label contigs as plasmid being far from the largest plasmid contig.

The separation beween ambiguous and chromosome contigs is much less clear, although the number of long ambiguous contigs is quite small.

The main question is about the large number of relatively short ambiguous contigs: how many true plasmid contigs do-we miss?

In terms of subsequent analysis, the impact of ambiguous contigs is also crucial: how should we handle such contigs if a plasmid prediction tool includes some in predicted plasmids.

We look at the impact of the chosen threshold **L** on the number of contigs in each class.

`Length thresholds from 100000 to 2000000 (bp) in multiples of 100000`

The choice of **L** doe not impact significantly the number of ambiguous contigs, that are the main problem.

Landscape of contigs distribution We look at all isolates separately to see how the hybrid contigs are classified.

The table below shows that a large number of sample have many ambiguous contigs (i.e. the large number of ambiguous contigs does not originate from a handful of isolates), and there are many samples with no plasmid contig and many ambiguous contigs.

	chromosome	ambiguous	plasmid
SRR14000614	2.0	80.0	1.0
SRR14024961	0.0	69.0	0.0
SRR13726590	2.0	64.0	3.0
SRR13727011	1.0	63.0	0.0
SRR13726582	1.0	49.0	2.0
SRR13999927	2.0	46.0	1.0
SRR13726551	1.0	44.0	0.0
SRR13999934	2.0	39.0	4.0
SRR13727008	4.0	29.0	1.0
SRR14000594	1.0	24.0	1.0
SRR13712363	3.0	22.0	1.0
SRR13725703	1.0	21.0	0.0
SRR13726554	1.0	20.0	1.0
SRR13712377	1.0	16.0	0.0
SRR13725727	1.0	15.0	1.0
SRR14000019	1.0	15.0	0.0
SRR13726529	1.0	13.0	1.0
SRR13726578	1.0	12.0	0.0
SRR13999918	1.0	11.0	0.0
SRR13726561	1.0	11.0	2.0
SRR13999995	1.0	11.0	0.0
SRR13725706	1.0	9.0	0.0

SRR14000592	3.0	9.0	2.0
SRR13712371	1.0	8.0	4.0
SRR13727002	1.0	8.0	2.0
SRR13999953	1.0	8.0	0.0
SRR13726569	1.0	7.0	1.0
SRR13712383	2.0	7.0	3.0
SRR13727000	1.0	7.0	1.0
SRR13726557	1.0	6.0	3.0
SRR13712375	1.0	6.0	5.0
SRR13712391	1.0	6.0	3.0
SRR13725714	1.0	6.0	0.0
SRR13727003	1.0	6.0	2.0
SRR13726564	1.0	5.0	0.0
SRR14000598	1.0	5.0	0.0
SRR13999996	1.0	4.0	1.0
SRR13999990	1.0	4.0	0.0
SRR13726512	1.0	4.0	2.0
SRR13712366	1.0	3.0	3.0
SRR13726589	1.0	3.0	1.0
SRR13712379	1.0	3.0	3.0
SRR13999994	1.0	3.0	1.0
SRR13712518	1.0	2.0	0.0
SRR14000004	1.0	2.0	0.0
SRR13726536	1.0	2.0	2.0
SRR13999940	1.0	2.0	1.0
SRR13712361	1.0	2.0	1.0
SRR13726594	1.0	1.0	2.0
SRR13726587	1.0	1.0	1.0
SRR13727034	1.0	1.0	1.0
SRR14000018	1.0	0.0	1.0
SRR13999937	1.0	0.0	1.0
SRR13725701	1.0	0.0	1.0
SRR13726504	1.0	0.0	1.0
SRR13727026	1.0	0.0	1.0
SRR13727025	1.0	0.0	1.0
SRR13999931	1.0	0.0	2.0
SRR13712384	1.0	0.0	0.0
SRR14000005	1.0	0.0	1.0

Summary We summarize before presenting a similar analysis for the *E. faecium* samples, that draws the same conclusion.

Despite using long reads, the assemblies we are provided with are characterized by a large number of short linear contigs. Our current labelling method, that defines the ground truth we rely on to evaluate the accuracy of plasmid prediction methods, labels such contigs as *ambiguous*. We need - either to define how ambiguous contigs are handled when we evaluate a method, - or to be mre precise in identifying in hybrid assemblies which contigs are trule plasmid contigs.

The table above shows that the issue is across the whole dataset and can not be fixed by discarding few samples. The fact there are only 24 isolates with a large circular chromosome corresponbding to a full assembly of the chromosome rules out that we can improve significantly the issue described above by refining the labellinbg method in such cases (in which we could consider all other contigs as plasmids).

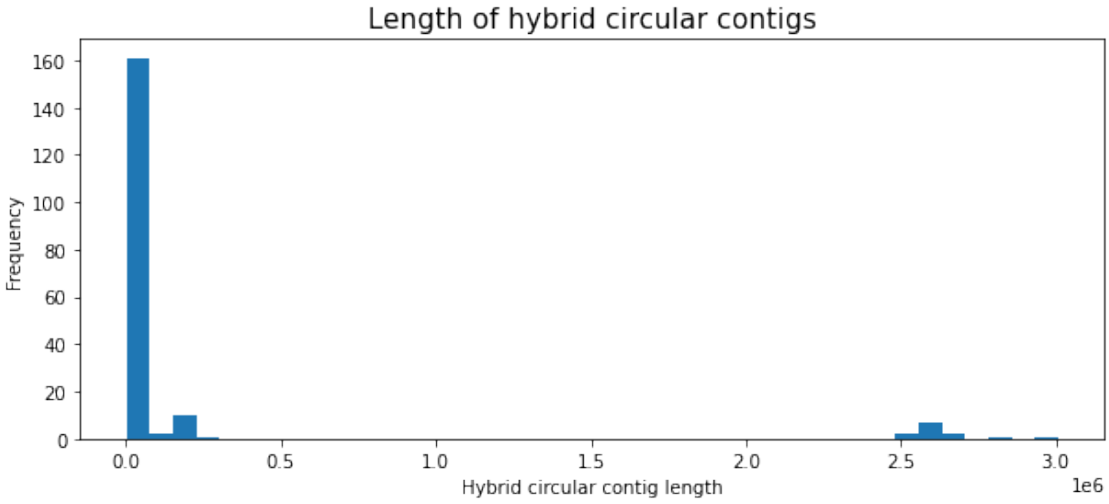
1.4.1 E. faecium

We present our analysis of the *E. faecium* isolates. (Contig refer to hybrid contigs.)

Statistics on isolates and contigs

```
The total number of E. faecium isolates is 71 and of contigs over all isolates
is 3538
The number of contigs labelled as plasmid over all isolates is 174
The number of samples with no contig labelled as plasmid is 12
```

Length distribution of circular contigs

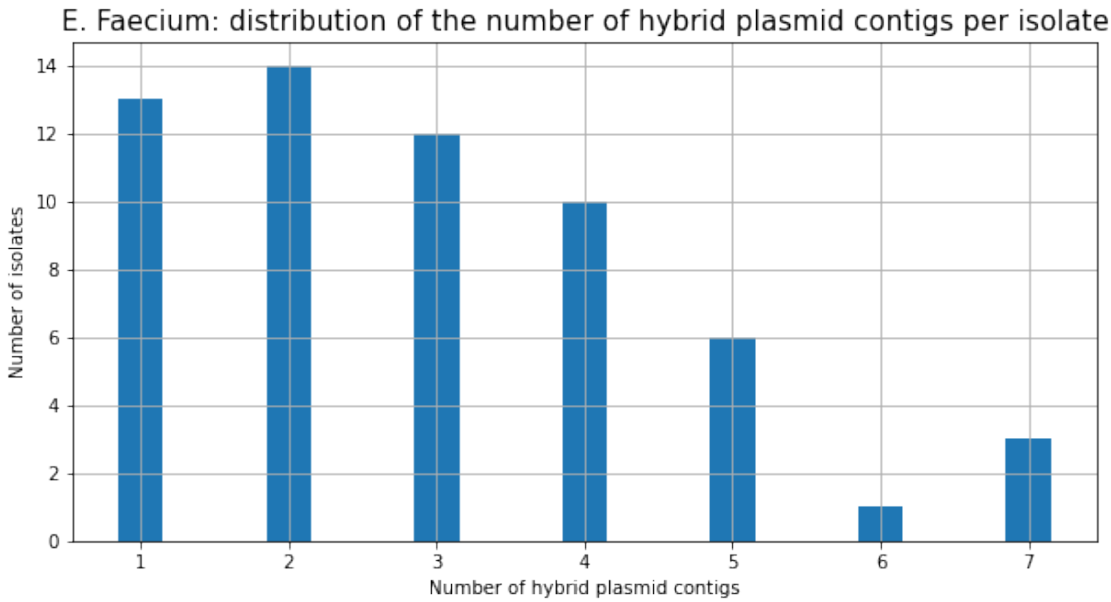


As with E. faecalis, the choice for L is left unclear. We choose a small value of L = 300000.

```
Over all E. faecium isolates there are 187 circular contigs
  13 are classified as chromosome
  174 are classified as plasmid
```

E. faecium shows the same trend as E. faecalis. The number of chromosome contigs is very small compared to the total number of isolates indicating that not all chromosomes are not fully assembled.

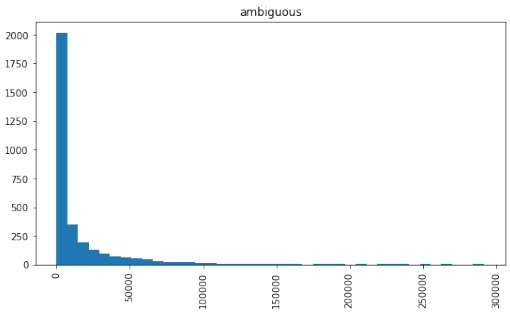
Hybrid contigs classified as plasmids

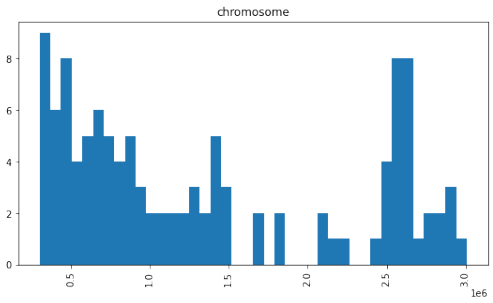


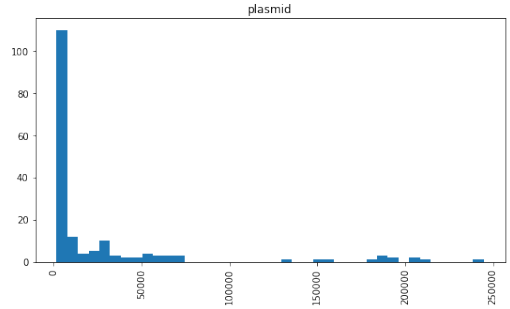
Distribution of contig lengths per label We now look at the three classes of contigs (plasmid, chromsome and ambiguous) for E. faecium isolates.

```
Number of chromosome contigs: 174
Number of plasmid contigs: 114
Number of ambiguous contigs: 3250
```

```
Hybrid contig length distribution per class
```

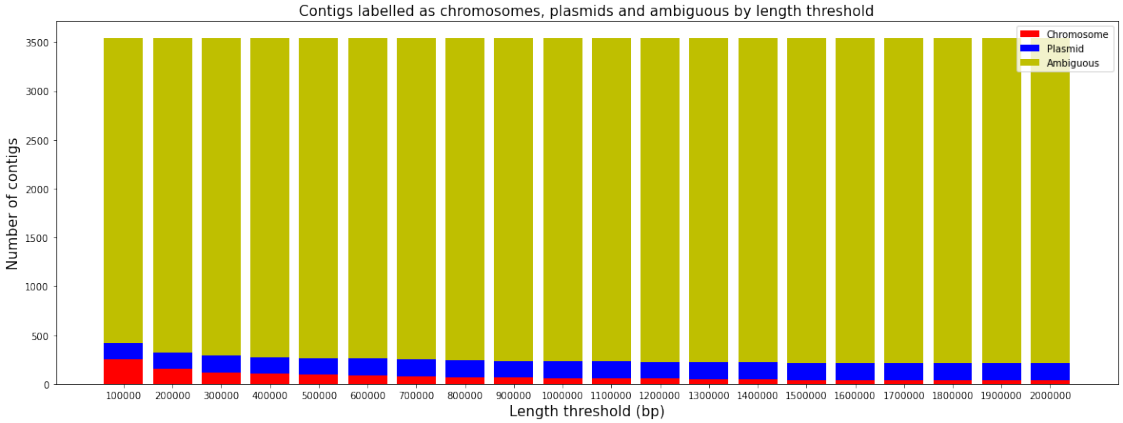






3424 out of 3538 contigs are under 300000 bp in length and 3250 of these contigs have been classified as ambiguous.

All of the ambiguous and plasmid contigs are shorter than 150000 bp. This indicates a strong separation between ambiguous/plasmid and chromosome contigs, with the chosen threshold (300000bp). However, it is unclear if ambiguous contigs are fragments of chromosomes or should be considered independent plasmids.



Landscape of contigs distribution

First 50 E. faecium samples

	chromosome	ambiguous	plasmid
SRR14022764	1.0	237.0	0.0
SRR14010969	0.0	184.0	4.0
SRR14011002	0.0	173.0	4.0
SRR14026515	0.0	170.0	2.0
SRR14010946	0.0	167.0	2.0
SRR14011041	0.0	166.0	3.0
SRR14010961	1.0	148.0	3.0
SRR14022770	0.0	143.0	0.0
SRR14022782	2.0	129.0	1.0
SRR14026532	0.0	128.0	0.0
SRR14022776	2.0	124.0	0.0
SRR14010955	1.0	101.0	5.0
SRR14026537	2.0	94.0	3.0
SRR14024950	2.0	90.0	0.0
SRR14026549	0.0	79.0	0.0
SRR14011022	4.0	69.0	4.0
SRR14010981	2.0	59.0	4.0
SRR14011043	3.0	53.0	3.0

SRR14024983	2.0	51.0	2.0
SRR14011035	4.0	49.0	5.0
SRR14010965	5.0	48.0	3.0
SRR14026517	2.0	42.0	5.0
SRR14010950	4.0	42.0	2.0
SRR14026545	3.0	41.0	0.0
SRR14022735	2.0	39.0	1.0
SRR14010988	4.0	34.0	7.0
SRR14010953	2.0	33.0	4.0
SRR14026530	2.0	32.0	4.0
SRR14026535	2.0	30.0	7.0
SRR14010933	3.0	29.0	4.0
SRR14026552	3.0	28.0	2.0
SRR14011003	4.0	27.0	5.0
SRR14010983	3.0	26.0	4.0
SRR14026514	3.0	26.0	2.0
SRR14024951	1.0	26.0	1.0
SRR14010971	1.0	23.0	6.0
SRR14022763	1.0	20.0	3.0
SRR14011020	2.0	19.0	2.0
SRR14022769	1.0	18.0	4.0
SRR14022761	1.0	17.0	0.0
SRR14011039	1.0	16.0	4.0
SRR14026541	1.0	15.0	2.0
SRR14026503	2.0	14.0	1.0
SRR14022777	1.0	14.0	1.0
SRR14022778	1.0	13.0	1.0
SRR14011001	1.0	11.0	1.0
SRR14024986	1.0	11.0	2.0
SRR14024997	3.0	11.0	1.0
SRR14024963	1.0	10.0	1.0
SRR14024990	1.0	10.0	2.0

Summary Overall similar to *E. faecalis* but for a more serious issue with the number of ambiguous contigs, and some cases where the assembly is so fragmented that there are actually no contig labelled as chromosome.